CS 7545: Machine Learning Theory **Fall 2021**

Lecture 7

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Clustering

The objective of clustering is to partition a set into dissimilar subsets $S_1, S_2, \ldots, S_k$ of similar elements (often measured relative to set "centers" $c_1, c_2, \ldots, c_k$). It is often represented as an optimization problem on a graph $G = (V, E)$ with distance metric $d(\cdot, \cdot)$. Some common objective functions are

- Diameter: $\min \max_{1 \le i \le k, x, y \in S_i} d(x, y)$.
- k-center: $\min \max_{x \in V, x \in S_i} d(x, c_i)$.
- k-median: $\min \sum_{x \in V, x \in S_i} d(x, c_i)$.
- k-means: $\min \sum_{x \in V, x \in S_i} d^2(x, c_i)$.

These problems are all **NP**-hard. So, a common setting is to assume that a ground-truth clustering exists, and we'd like to (probably approximately) recover it.

2 k-center Approximation

We will begin by analyzing a greedy approximation algorithm for the k-center problem. We start with any point $C = \{c_1\}$, then add the furthest point from any element of C to C, and repeat $k-1$ times.

Theorem 2.1: k-center Approximation

The greedy k-center algorithm gives a 2-approximation to the optimal solution.

Proof: Suppose $OPT = r$, and assume for a contradiction that the cost of the greedy algorithm solution C is greater than $2r$. Then, there must exist a point c_{k+1} at some distance greater than $2r$ from any element of C. Because we made greedy choices, this implies $d_{i \neq j}(c_i, c_j) > 2r$. Since the optimal solution uses k clusters, it must put at least two of $c_1, c_2, \ldots, c_{k+1}$ in the same cluster. But then one of those points is distance greater than r from its center (by the triangle inequality), so $OPT > r$, a contradiction. ■

3 Spectral Clustering

We will now analyze a general algorithm which is applicable for a wide variety of different clustering problems, as long as there is a well-separated ground-truth clustering. Suppose $\mathbf{A} \in \mathbb{R}^{n \times d}$ is the dataset and each row is a datapoint. We'd like to find $\mathbf{C} \in \mathbb{R}^{n \times d}$ with k distinct rows corresponding to cluster centers (it will happen that each row in $\mathbf{C}$ is the center for the corresponding row of $\mathbf{A}$, but even if not, it is easy to compute).

Recall that the matrix 2-norm is the largest singular value (*i.e.*, the maximum action in any direction) and the matrix Frobenius norm is the Euclidean norm of the elements, or equivalently the Euclidean norm of the singular values. In the k-means setting, our objective is

$$\min_{\mathbf{C}} \|\mathbf{A} - \mathbf{C}\|_F^2 = \sum_{i=1}^{n} \|\boldsymbol{A}_i - \boldsymbol{C}_i\|_2^2. \tag{1}$$

Definition 3.1: Average Intracluster Variance

The average intracluster variance is

$$\sigma^2(\mathbf{C}) = \frac{1}{n}\|\mathbf{A} - \mathbf{C}\|_2^2. \tag{2}$$

This roughly measures the average maximum "spread", or variance, within each cluster.

The algorithm is as follows:

1. Project $\mathbf{A}$ to its SVD subspace $\mathbf{A}_k$.
2. Pick a random row of $\mathbf{A}_k$ and include all points within distnace $D = \frac{6k}{\epsilon}\sigma(\mathbf{C})$ in its cluster. Remove these points afterwards.
3. Repeat until we have k clusters.

Theorem 3.1: Spectral Clustering

Suppose there exists $\mathbf{C}$ with $\|\boldsymbol{C}_i - \boldsymbol{C}_j\| \geq \frac{15k}{\epsilon}\sigma(\mathbf{C})$ and each cluster contains at least ϵn points. Then, the spectral clustering algorithm finds a clustering which differs from $\mathbf{C}$ in at most $\epsilon^2 n$ points. Note that the guarantee is dimension independent.

We will first show a lemma which enables us to bound $\|\mathbf{A}_k - \mathbf{C}\|_F$. Typically, $\|\mathbf{A}\|_F^2 \leq \operatorname{rank}(\mathbf{A}) \cdot \|\mathbf{A}\|_2^2$, but spectral projection allows us to be linear in k instead of d.

Lemma 3.1: Frobenius Norm of Spectral Projection

Suppose $\mathbf{A}_k$ is the best rank k approximation of $\mathbf{A}$, that is,

$$\mathbf{A}_k = \underset{\mathbf{D}:\mathrm{rank}(\mathbf{D})\leq k}{\mathrm{argmin}} \|\mathbf{A}-\mathbf{D}\|_2. \tag{3}$$

Then, for $\mathbf{C}$ of rank k and any $\mathbf{A}$ we have

$$\|\mathbf{A}_k - \mathbf{C}\|_F^2 \leq 8k\|\mathbf{A}-\mathbf{C}\|_2^2 = 8k\sigma^2(\mathbf{C})n. \tag{4}$$

Proof: $\mathbf{A}_k - \mathbf{C}$ has rank at most $2k$ so

$$\|\mathbf{A}_k - \mathbf{C}\|_F^2 \leq 2k\|\mathbf{A}_k-\mathbf{C}\|_2^2. \tag{5}$$

Because $\mathbf{A}_k$ is a 2-norm minimizer and $\mathbf{C}$ was a candidate in this optimization,

$$\begin{aligned}\|\mathbf{A}_k - \mathbf{C}\|_2 &\leq \|\mathbf{A}_k - \mathbf{A}\|_2 + \|\mathbf{A}-\mathbf{C}\|_2, &(6)\\ &\leq 2\|\mathbf{A}-\mathbf{C}\|_2. &(7)\end{aligned}$$

Combining the two equations yields the result. ■